Linear mixed models for replication data to efficiently allow for covariate measurement error

Jonathan W. Bartlett, Bianca L. De Stavola and Chris Frost

Department of Medical Statistics, London School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT

Abstract

It is well known that measurement error in the covariates of regression models generally causes bias in parameter estimates. Correction for such biases requires information concerning the measurement error, which is often in the form of internal validation or replication data. Regression calibration (RC) is a popular approach to correct for covariate measurement error, which involves predicting the true covariate using error-prone measurements. Likelihood methods have previously been proposed as an alternative approach to estimate the parameters in models affected by measurement error, but have been relatively infrequently employed in medical statistics and epidemiology, partly because of computational complexity and concerns regarding robustness to distributional assumptions. We show how a standard random-intercepts model can be used to obtain maximum likelihood (ML) estimates when the outcome model is linear or logistic regression under certain normality assumptions, when internal error-prone replicate measurements are available. Through simulations we show that for linear regression, ML gives more efficient estimates than RC, although the gain is typically small. Furthermore, we show that RC and ML estimates remain consistent even when the normality assumptions are violated. For logistic regression, our implementation of ML is consistent if the true covariate is conditionally normal given the outcome, in contrast to RC. In simulations, this ML estimator showed less bias in situations where RC gives non-negligible biases. Our proposal makes the ML approach to dealing with covariate measurement error more accessible to researchers, which we hope will improve its viability as a useful alternative to methods such as RC.

1 Introduction

Standard statistical regression models assume that the covariates or explanatory variables in the regression model are observed exactly. While this is true for certain variables, observed values of variables can often be considered error-prone measurements of the true covariates in which we are primarily interested. It is well known that in general, ignoring error in covariates results in biased estimates of the parameters of the model relating the outcome to the true covariates [1, 2].

Many different methods have been developed for correcting the bias in such naive estimates, with the aim of estimating the association between the outcome and true covariates of interest using the observed error-prone measurements. Usually, such correction methods require information about the measurement error, often in the form of replication or validation datasets, whereby independent replicate error-prone measurements or the true covariate are observed for a subset of

subjects. Furthermore, depending on the correction method employed, one must usually also make assumptions about the true covariate and the measurement error, which would not be necessary had the true covariate been observed without error.

Regression calibration (RC) is arguably the method most commonly used in epidemiology to allow for covariate measurement error [2]. For a linear regression outcome model, regression calibration gives consistent estimates of the parameters of interest. For non-linear outcome models such as logistic regression, regression calibration has been shown to be approximately consistent, although non-trivial biases can occur if the covariate effects are strong or the measurement error is large [3].

Maximum likelihood (ML) has also been advocated for allowance for covariate measurement error [4, 5], and it is an established technique in common use for fitting structural equation models where latent variables represent the true covariate [6]. To use parametric likelihood, one must specify a parametric model for the observed data, i.e. the outcome and the error-prone measurements of the covariate. A model for the 'full' or 'complete' data can be specified as the product of the outcome model, the measurement error model, and a model for the true covariate, so long as covariate measurement error is assumed to be non-differential [7]. The likelihood function for the observed data is found by marginalizing the distribution of the 'full' data over the distribution of the true covariate, which is unobserved or 'latent'. For a correctly specified parametric model, ML is consistent and asymptotically efficient, and so ought to be considered as an alternative to simpler methods such as regression calibration.

A number of factors have probably contributed to parametric likelihood methods not having been used more often for covariate measurement error correction. First, commands to fit models via ML which allow for covariate measurement error are not universally available in statistical software packages. Notable exceptions include the CME wrapper for GLLAMM in Stata [5], Mplus [8], and proc NLMIXED in SAS, which can be adapted to fit joint models [9]. In contrast, regression calibration can be implemented relatively easily using standard statistical packages. Second, maximizing the likelihood function for the observed data can be computationally burdensome. Unless joint normality is assumed for the outcome, true covariate and measurement errors, the density function of the observed data generally involves an intractable integral over the distribution of the unobserved true covariate. Numerical integration methods such as Gaussian quadrature have being successfully employed to deal with this, although estimation can often be slow, especially with multiple true covariates. Third, researchers may be concerned that the strong assumptions made by parametric ML may not be valid, which may violate the validity of resulting estimates and inferences.

In this paper we aim to address both the issue of computational complexity and robustness for the ML approach when independent replicate error-prone measurements are available. We show that when the outcome model is linear regression, ML estimates for the joint model can be readily found by fitting a standard random-intercepts linear mixed model, routines for which are widely available in statistical packages. We show that this approach can also be used to find ML estimates when the outcome model is logistic regression and the true covariate is conditionally normal, given the outcome. We also show how Wald-type confidence intervals can be obtained using the estimates and standard errors reported by the fitted linear mixed model, and for logistic regression, how Fieller's theorem can be used to construct confidence intervals (Appendix A). For both types of outcome models, we use simulations to investigate the finite sample performance of RC and ML, using characteristics which may be typical of epidemiological studies. We also use simulations to explore robustness of RC and ML to some of their parametric assumptions. For linear regression, we show that the ML estimator remains consistent even when the normality

assumptions upon which it is derived do not hold. We introduce our approach in the context of simple linear and logistic regression with a single covariate which is measured with error, and in Appendix B extend the approach to the case of multiple covariates, some of which may be measured without error.

2 Linear regression

2.1 Model specification

We assume that data are available for $i = 1, .., n$ independent subjects. We suppose that an outcome Y_i follows a linear regression model conditional on the covariate of interest, X_i:

$$Y_i = \alpha + \beta X_i + \epsilon_i \tag{1}$$

where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ and is independent of X_i. Furthermore, we assume that $X_i \sim N(\mu_X, \sigma_X^2)$. The outcome Y_i is observed, but X_i is not. Instead, for the ith subject, N_i error-prone measurements, $\mathbf{W}_i = (W_{i1}, .., W_{iN_i})^T$ of X_i are available. In general, the number of error-prone measurements for subject i, N_i, is a random variable with some discrete distribution taking positive integer values. A common study design is one in which a randomly selected proportion π of subjects have $N_i = 2$ error-prone measurements of X_i, with the remaining subjects having $N_i = 1$. This corresponds to N_i taking value 2 with probability π and taking value 1 with probability $1 - \pi$. We condition on the realized values of N_i throughout, on the assumption that N_i is independent of all other variables and therefore contains no information regarding the parameters of interest. We note that sometimes N_i may be fixed by design, e.g. $N_i = 2$ for all subjects.

We assume a classical measurement error model, so that:

$$W_{ij} = X_i + U_{ij} \tag{2}$$

where $U_{ij} \sim N(0, \sigma_U^2)$ for $j = 1, .., N_i$. The measurement errors U_{ij} are assumed independent of each other, and of X_i. We further assume that the measurement errors are non-differential, so that Y_i and $\mathbf{W}_i$ are independent conditional on X_i. This is equivalent to U_{ij} being independent of the residual errors ϵ_i. Together these assumptions define a joint model for the observed data and the unobserved true covariate X_i. Figure 1 shows diagramatically the relationship between the observed variables $\mathbf{W}_i$ and Y_i, and the unobserved variable X_i when $N_i = 2$.

The ratio of the variance of the true covariate to that of the error prone measurements:

$$\lambda = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2}$$

is known as the reliability ratio. Ignoring covariate measurement error and fitting the regression of Y_i on a single measurement W_{i1} leads to an unbiased estimate of $\lambda\beta$ [1].

2.2 Regression calibration

Regression calibration (RC) involves calculating $E(X_i|\mathbf{W}_i)$ (or an estimate of this) for each subject, and then fitting the linear regression of Y_i on these values. Under the assumed model, $\mathbf{W}_i$ follows a one-way random-intercepts model, with random effect variance σ_X^2 and within-subject variance σ_U^2. The components of variance can be estimated using the ANOVA estimators, or under the normality

assumptions previously described, using either ML or restricted ML (REML) [11], giving estimates of σ_X^2, σ_U^2 and μ_X. Under this assumed model, $\mathbf{W}_i$ and X_i are jointly normal, and from standard results for multivariate normal distributions (e.g. Appendix S3 of [11]) we have:

$$E(X_i|\mathbf{W}_i) = \mu_X + \frac{\sigma_X^2}{\sigma_X^2 + \frac{\sigma_U^2}{N_i}} \left(\frac{\sum_{j=1}^{N_i} W_{ij}}{N_i} - \mu_X \right). \tag{3}$$

Regression calibration then proceeds by fitting the regression of Y_i on $\hat{E}(X_i|\mathbf{W}_i)$, which is found by substituting the estimates of σ_X^2, σ_U^2 and μ_X into equation (3). The slope estimate from this regression model, $\hat{\beta}_{\text{RC}}$, is then an estimate of β. When the outcome model is linear regression, $\hat{\beta}_{\text{RC}}$ is a consistent estimator of β so long as the parameters needed to calculate $E(X_i|\mathbf{W}_i)$ are consistently estimated as $n \to \infty$. This is satisfied when $n \times P(N_i > 1) \to \infty$ as $n \to \infty$, which ensures that the number of subjects with two or more error-prone measurements tends to infinity.

We note that naive standard errors and confidence intervals for $\hat{\beta}_{\text{RC}}$ obtained from fitting the model for Y_i with $\hat{E}(X_i|\mathbf{W}_i)$ ignore uncertainty in the parameters involved in $E(X_i|\mathbf{W}_i)$. One approach to valid inference is to use bootstrapping (see Appendix A.9.5 of [2]).

2.3 Maximum likelihood

To find the ML estimates of the joint model, we maximize the likelihood function given the observed data, Y_i and $\mathbf{W}_i$. Our strategy is to factorise the log likelihood function as:

$$\begin{aligned} l(\boldsymbol{\theta}|Y_i, \mathbf{W}_i) &= \log(f(Y_i, \mathbf{W}_i|\boldsymbol{\theta})) \\ &= \log(f(Y_i|\boldsymbol{\theta})) + \log(f(\mathbf{W}_i|Y_i, \boldsymbol{\theta})) \end{aligned} \tag{4}$$

where $\boldsymbol{\theta} = (\alpha, \beta, \sigma_\epsilon^2, \mu_X, \sigma_X^2, \sigma_U^2)$ is the vector of model parameters. Since X_i and ϵ_i are independent and normal, it follows that Y_i is marginally normal with mean and variance:

$$\mu_Y = \alpha + \beta\mu_X \tag{5}$$

$$\sigma_Y^2 = \beta^2\sigma_X^2 + \sigma_\epsilon^2. \tag{6}$$

Furthermore, since X_i and Y_i are jointly normal, X_i given Y_i is also normal. We can therefore write:

$$X_i = \gamma_0 + \gamma_Y Y_i + b_i \tag{7}$$

where $b_i \sim N(0, \sigma_{X|Y}^2)$ is an independent normally distributed residual and:

$$\gamma_0 = \mu_X - \frac{\beta\sigma_X^2(\alpha + \beta\mu_X)}{\beta^2\sigma_X^2 + \sigma_\epsilon^2} \tag{8}$$

$$\gamma_Y = \frac{\beta\sigma_X^2}{\beta^2\sigma_X^2 + \sigma_\epsilon^2} \tag{9}$$

$$\sigma_{X|Y}^2 = \sigma_X^2 - \frac{\beta^2\sigma_X^4}{\beta^2\sigma_X^2 + \sigma_\epsilon^2} \tag{10}$$

follow from standard results for multivariate normal distributions (e.g. Appendix S3 of [11]). Then since $W_{ij} = X_i + U_{ij}$, substituting for X_i using equation (7) we have

$$W_{ij} = \gamma_0 + \gamma_Y Y_i + b_i + U_{ij}$$

where $U_{ij} \sim N(0, \sigma_U^2)$ is independent of b_i. This shows that W_{ij} given Y_i follows a random-intercepts model, with random-intercepts variance $\sigma^2_{X|Y}$, within-subject variance σ_U^2, and a fixed effect of Y_i.

The parameter vector $\boldsymbol{\phi} = (\mu_Y, \sigma_Y^2, \gamma_0, \gamma_Y, \sigma^2_{X|Y}, \sigma_U^2)$ is a one-to-one function of the original model parameter vector $\boldsymbol{\theta} = (\alpha, \beta, \sigma_\epsilon^2, \mu_X, \sigma_X^2, \sigma_U^2)$. Furthermore, in terms of the new parametrization $\boldsymbol{\phi}$, $f(Y_i)$ and $f(\mathbf{W}_i|Y_i)$ share no parameters. Since the two subsets of parameters are variationally independent, it follows that the ML estimate of $\boldsymbol{\phi}$ can be obtained by maximizing the two likelihood components separately [12].

In the first part of the log likelihood (equation (4)), $\log(f(Y_i|\boldsymbol{\phi}))$, the ML estimates of μ_Y and σ_Y^2 are simply the sample mean and sample variance with n in the denominator. The second part of the log likelihood, $\log(f(\mathbf{W}_i|Y_i, \boldsymbol{\phi}))$, is that of a random-intercepts model, and so standard software can be used to fit a random-intercepts model for $\mathbf{W}_i$ given Y_i, using either ML or restricted ML (REML), giving estimates $\hat{\gamma}_0$, $\hat{\gamma}_Y$, $\hat{\sigma}^2_{X|Y}$, $\hat{\sigma}^2_U$.

To obtain the ML estimate of β, we solve for β using equations (6), (9) and (10), which gives:

$$\beta = \frac{\gamma_Y \sigma_Y^2}{\sigma^2_{X|Y} + \gamma_Y^2 \sigma_Y^2}. \tag{11}$$

The ML estimate of β can thus be calculated as:

$$\hat{\beta}_{\mathrm{ML}} = \frac{\hat{\gamma}_Y \hat{\sigma}_Y^2}{\hat{\sigma}^2_{X|Y} + \hat{\gamma}_Y^2 \hat{\sigma}_Y^2}. \tag{12}$$

Like $\hat{\beta}_{\mathrm{RC}}$, $\hat{\beta}_{\mathrm{ML}}$ is consistent if as $n \to \infty$, $n \times P(N_i > 1) \to \infty$. In Appendix A.1 we show how Wald type confidence intervals for $\hat{\beta}_{\mathrm{ML}}$ can be constructed, using the estimates and corresponding standard errors for $\hat{\gamma}_Y$, $\hat{\sigma}^2_{X|Y}$ and $\hat{\sigma}_Y^2$. In Appendix B.1 we show how the approach can be extended to include multivariate $\mathbf{X}_i$ and error free covariates $\mathbf{Z}_i$.

2.4 Simulations

In Table 1 we give simulation results comparing RC with ML, simulated under the model described previously. For all scenarios, without loss of generality, we set $\beta = 1$ and $\sigma_X^2 = 1$. We varied the value of σ_ϵ^2 to consider either weak, moderate or strong associations between Y_i and X_i, corresponding to a correlation between Y_i and X_i of 0.2, 0.5 and 0.8. Values for the measurement error variance σ_U^2 of 1/2, 1 and 2 were used, corresponding to reliability ratios λ of 2/3, 1/2 and 1/3 respectively. The results for each scenario are based on 10,000 simulations, with n=5,000 subjects, 500 of whom have two error-prone measurements, and the remaining 4,500 have just one. The simulations were performed using R 2.8.0 [13], using the lmer command of the lme4 package to fit the random-intercepts models. Wald-type confidence intervals were calculated as described in Appendix A.1, using the reported standard error for $\hat{\gamma}_Y$ from lmer. We calculated the standard error for $\hat{\sigma}^2_{X|Y}$ using the appropriate element of the inverted observed information matrix.

RC and ML are consistent for a linear regression outcome model under the conditions previously described, and our simulations showed little bias for the sample size and scenarios considered (Table 1). The variability of RC was very similar to ML - ML only being appreciably more efficient for scenario 9, when the reliability ratio was 1/3 and the association between X_i and Y_i very strong. Indeed for most scenarios, the RC and ML estimate for a given dataset were very similar. For example, for scenario 1, the standard deviation of the difference between them was 0.002.

The empirical coverage of the Wald confidence intervals for $\hat{\beta}_{ML}$ were close to their nominal 95% level for all scenarios. However, as λ decreased and the correlation between Y_i and X_i increased, the coverage rates of the one-sided intervals using the lower limit of the 95% interval became closer to 100% while the coverage of the one-sided interval using the upper limit became closer to 95%, rather than their 97.5% nominal coverage rates. This was a symptom of the fact that the sampling distribution of $\hat{\beta}_{ML}$ was positively skewed. The result was that the two-sided Wald intervals had close to the correct coverage level.

2.5 Relationship between RC, ML and method of moments

Our simulations show that for the setup used, which is typical of many epidemiological studies, $\hat{\beta}_{RC}$ and $\hat{\beta}_{ML}$ are not identical, although they usually give very similar point estimates. In the special case in which $N_i = N_\bullet$ for some $N_\bullet > 1$ for all subjects, RC and ML are asymptotically identical, i.e. $\hat{\beta}_{RC} = \hat{\beta}_{ML}$ with a probability that tends to 1 as $n \to \infty$ and $n \times P(N_i > 1) \to \infty$. In this case, $\hat{\beta}_{RC}$ is also identical to a method of moments estimate of β, which is obtained by dividing the naive estimate from regressing Y_i on the mean of the k measurements $\overline{W}_i$, by:

$$\frac{\hat{\sigma}_X^2}{\hat{\sigma}_X^2 + \frac{\hat{\sigma}_U^2}{N_\bullet}}, \tag{13}$$

where the estimates of $\hat{\sigma}_X^2$ and $\hat{\sigma}_U^2$ are the same ones used by RC to calculate $\hat{E}(X_i|\mathbf{W}_i)$.

More typically, the number of replicates N_i is different for different subjects. In this case, the residual variance of Y_i given $E(X_i|\mathbf{W}_i)$ (and $\hat{E}(X_i|\mathbf{W}_i)$) differs for subjects with different values of N_i. This heterogeneity in residual variance is ignored by the OLS estimator of β. This means that the implementation of RC we have considered is inefficient compared to ML. The residual variance of the regression of Y_i on $E(X_i|\mathbf{W}_i)$ for subject i can be shown to be equal to:

$$\sigma_\epsilon^2 + \beta^2\sigma_X^2\left(1 - \frac{\sigma_X^2}{\sigma_X^2 + \frac{\sigma_U^2}{N_i}}\right). \tag{14}$$

This suggests that the inefficiency of RC is likely to be small if either $\beta^2\sigma_X^2$ is small relative to σ_ϵ^2, i.e. when X_i explains only a small proportion of the variation in Y_i, or if σ_U^2 is small relative to σ_X^2. Our simulation results in Tables 1 and 2 appear to confirm this, with efficiency differences between RC and ML only apparent when the correlation between Y_i and X_i was large and the reliability ratio λ was small.

2.6 Robustness

A natural concern for methods which adjust for covariate measurement error is robustness to any additional assumptions which are made, in particular those which would not be necessary if the true covariate X_i were observed. Without normality assumptions, the expression given in equation (3) for $E(X_i|\mathbf{W}_i)$ is the best (in the sense of minimizing mean squared error of prediction) linear prediction of X_i given $\mathbf{W}_i$ (see Appendix 4 of [2]). Furthermore the mean μ_X and variances σ_X^2 and σ_U^2 involved in $E(X_i|\mathbf{W}_i)$ are estimated consistently by the ML estimator (derived under normality assumptions) even if X_i or the U_{ij} are not normally distributed [18]. Lastly, for a linear regression outcome model, unbiased estimates of β are obtained by fitting the model using the best linear prediction of X_i given $\mathbf{W}_i$ [17]. It thus follows that the RC estimator $\hat{\beta}_{RC}$ we have described is

consistent even if X_i or U_{ij} are not normally distributed. We also note that normality of the errors ϵ_i is not needed, since the consistency of the OLS estimators does not rely on such an assumption.

When σ_U^2 is known and $N_i = 1$ for all subjects, the ML estimator based on normality ('normal model ML') of X_i is consistent even if X_i is not normal [1]. It is therefore also of interest whether the ML approach based on normality assumptions, as described in Section 2.3, in which σ_U^2 is simultaneously estimated with the other model parameters, remains consistent under deviations from the normality assumptions. The assumed model includes assumptions of normality for X_i, U_{ij} and ϵ_i. If X_i and ϵ_i are not both normally distributed, the linear mixed model for $\mathbf{W}_i$ given Y_i is in general misspecified in a number of ways. First, the random effects b_i are not normally distributed. Second, the variance of these random effects may vary as a function of Y_i. Lastly, the conditional mean function $E(X_i|Y_i)$ may be a more complicated function of Y_i, and so the fixed effects structure of the mixed model may be misspecified. In Appendix C, we show that the likelihood score equations for the linear mixed model have zero expectation without requiring any normality assumptions. It then follows from the theory of estimating equations (see for example Appendix A6 of [2]) that the ML estimators solving these equations remain consistent without requiring normality of U_{ij}, X_i or ϵ_i. Similarly, the ML estimator of variance for σ_Y^2 is consistent regardless of the marginal distribution of Y_i. It thus follows that $\hat{\beta}_{ML}$ remains consistent even if some or all of the normality assumptions are violated.

To investigate the finite-sample performance of RC and normal model ML under non-normal X_i, we performed simulations in which X_i was log-normally distributed. If $A \sim N(0, \sigma^2)$, then $\mathrm{Var}(e^A) = e^{\sigma^2}(e^{\sigma^2} - 1)$ [14]. We therefore simulated X_i by exponentiating a random draw from $N(0, 0.481)$, so that $\mathrm{Var}(X_i) = 1$. The resulting distribution of X_i thus had skew of 2.84. Under this data generating model, the conditional distribution of X_i given Y_i is not normal, the conditional mean depends on higher order terms than merely Y_i, and the conditional variance of b_i, the best linear prediction residual, also varies with Y_i.

Table 2 shows the results of the simulations, with 10,000 simulations used for each scenario. As before, there was little evidence of bias in RC or normal model ML based on joint normality. The variability of both RC and normal model ML were similar to when X_i was normally distributed for a reliability ratio of 2/3 or 1/2. For a reliability ratio of 1/3, both RC and normal model ML were more variable than they were for normal X_i. The efficiency advantage of normal model ML over RC was greater compared to that when X_i was normally distributed. For the simulation results of Table 2 we used the same Wald confidence intervals which are derived on the basis of X_i being normally distributed, and so under non-normality we would expect the coverage to deviate from the nominal level. However, in our simulations the coverage of these confidence intervals was reasonable, only being slightly below the nominal 95% level, although the one-sided coverage rates were incorrect as described previously.

3 Logistic regression

We now suppose that Y_i is a binary random variable, which given a covariate X_i, follows a logistic regression model:

$$P(Y_i = 1|X_i) = \frac{\exp(\alpha + \beta X_i)}{1 + \exp(\alpha + \beta X_i)}. \tag{15}$$

As before, we assume that X_i is not observed, but that N_i error-prone measurements are available, as described earlier.

3.1 Regression calibration

Regression calibration is implemented in the same way as when the outcome model is linear regression (unless subjects are sampled on the basis of Y_i, i.e. case-control sampling). For logistic regression however, regression calibration results only in approximately consistent estimates of β.

3.2 Maximum likelihood assuming conditional normality

Analogous to the linear regression case, our strategy is to fit a linear mixed model for $\mathbf{W}_i$ given Y_i. We assume that $X_i|Y_i \sim N(\gamma_0 + \gamma_Y Y_i, \sigma^2_{X|Y})$. The assumption that X_i is normal given Y_i implies a logistic regression for Y_i given X_i [19], with:

$$\beta = \frac{\gamma_Y}{\sigma^2_{X|Y}}. \tag{16}$$

If the measurement errors U_{ij} are normal, it follows that W_{ij} given Y_i is normal with:

$$W_{ij} = \gamma_0 + \gamma_Y Y_i + b_i + U_{ij}.$$

Thus as with a linear regression outcome model, W_{ij} follows a random-intercepts model given Y_i, which can be fitted using standard software.

The 'conditional normal ML' estimate of β is then given by:

$$\hat{\beta}_{\text{ML}} = \frac{\hat{\gamma}_Y}{\hat{\sigma}^2_{X|Y}}. \tag{17}$$

In contrast to the case of a linear regression outcome model, under this model, β is a function only of parameters in the conditional distribution of $X_i|Y_i$. As for linear regression, an approximate 95% Wald confidence interval can be constructed for $\hat{\beta}_{\text{ML}}$. Since $\hat{\gamma}_Y$ and $\hat{\sigma}^2_{X|Y}$ are asymptotically normal and uncorrelated, we can also construct a 95% confidence interval for $\hat{\beta}_{ML}$ using Fieller's theorem (see Appendix A.2). In Appendix B.2 we show when our approach can be extended to accommodate multivariate $\mathbf{X}_i$ and error-free covariates $\mathbf{Z}_i$.

3.3 Simulations

We first compared the performance of RC with conditional normal ML under conditions for which our approach is indeed the correctly specified ML estimator, i.e. when $X_i|Y_i$ is normally distributed with constant variance. We considered both a rare ($P(Y_i = 1) = 0.1$) and a common outcome ($P(Y_i = 1) = 0.5$), a weak effect of X_i on Y_i ($\beta = 0.1$, corresponding to a standardized odds ratio (OR) of 1.11) and a relatively large effect ($\beta = 1$, corresponding to an OR of 2.72). As before, we considered reliability ratios of 2/3, 1/2 and 1/3.

The simulation results of Table 3 show that RC had little bias when the effect size was small. For a large effect size, there was a downward bias in RC, which was increased for larger $P(Y_i = 1)$. As expected, conditional normal ML had negligible bias for all scenarios. In general, RC was less variable than conditional normal ML, with the difference increasing with smaller values of the reliability ratio, larger effect sizes, or a more common outcome. Despite the bias of RC in these situations, the root mean square error (RMSE) of RC was less than conditional normal ML, except for scenarios 10 and 11. With increasing sample sizes, bias would dominate the RMSEs, so that conditional normal ML would have smaller RMSE than RC. Both the Wald-type confidence

intervals and those based on Fieller's theorem had coverage levels close to their nominal 95% level. However, while the one-sided coverage rates of the Wald intervals differed from 97.5% when λ decreased and β increased, the one-sided intervals based on Fieller's theorem had coverage close to their nominal 97.5% rate.

3.4 Robustness

The usual parametric model used in the context of covariate measurement error assumes marginal normality for X_i and normally distributed measurement error. If $X_i \sim N(\mu_X, \sigma_X^2)$, then unless $\beta = 0$, the conditional distribution of X_i given Y_i is not normal. However, as noted by Freedman et al [15, 16], if either the outcome Y_i is rare or the effect of X_i on Y_i is small, the distribution of X_i will be approximately both marginally and conditionally normal given Y_i. We therefore examined the robustness of our previously described conditional normal ML approach when X_i is marginally normal as opposed to conditionally normal given Y_i. With X_i marginally normal, our conditional normal ML estimator does not give the maximum likelihood estimate for the true data generating model. The likelihood function for the joint model with X_i marginally normal involves an intractable integral, which must be approximated either by analytical quadrature techniques or numerically via simulation.

With X_i marginally normal and Y_i given X_i logistic, the conditional distribution of X_i given Y_i is not normal (unless $\beta = 0$), and furthermore the conditional variance $\mathrm{Var}(X_i|Y_i)$ in general differs depending on whether $Y_i = 0$ or $Y_i = 1$. For the reasons outlined earlier for robustness in the case of linear regression, if we fit the linear mixed model for $\mathbf{W}_i$ given Y_i and $\mathrm{Var}(X_i|Y_i)$ depends on Y_i, the ML estimator of $\sigma^2_{X|Y}$ is a consistent estimator of the mean of the conditional variances $\mathrm{Var}(X_i|Y_i)$ over the distribution of Y_i, which here is equal to:

$$P(Y_i = 1)\mathrm{Var}(X_i|Y_i = 1) + P(Y_i = 0)\mathrm{Var}(X_i|Y_i = 0).$$

The bias in the 'conditional normal ML' estimator of β when X_i is marginally normal thus reduces to the bias of the method of normal discriminant analysis when its assumption of conditional normality with constant variance are violated.

Table 4 shows the results of simulations in which X_i was generated as marginally normal. As with the simulations with $X_i|Y_i$ normal, we see that the bias of RC is small except for when there is a large effect of X_i on Y_i. Despite the fact that the conditional normal MLE $\hat{\beta}_{ML}$ is not the MLE for the true data generating model in these simulations, it had little bias in the scenarios considered, presumably because normal discriminant analysis is relatively robust to the assumption of conditional normality [20]. Also of interest was that the coverage of the confidence intervals were very similar to when X_i was conditionally normal given Y_i.

4 Discussion

We have proposed a new approach for obtaining maximum likelihood estimates of covariate effects in linear and logistic regression in which the covariates are measured with error. We have restricted our attention to situations in which the true covariate cannot be observed, but internal replicate error-prone measurements are available for a subset of subjects. By using standard linear mixed models to model the error-prone measurements conditional on the outcome, our approach efficiently accommodates varying numbers of error-prone measurements between subjects.

For a linear regression outcome model, ML had smaller sampling variability than RC in simulations, although the differences were small for most scenarios. The greatest efficiency advantage for ML occurred when measurement errors were large and when the association between true covariate and outcome was also large. Both our implementation of RC and the normal model ML estimators are predicated on normality of the true covariate, the measurement errors, and the outcome regression error, but remain consistent even when some or all of these are not normally distributed. Our simulations confirm this property in the case of a log-normally distributed true covariate. Bootstrapping or sandwich type estimators of variance could be used to construct confidence intervals for ML estimates when the normality assumptions are in doubt, although in our simulations confidence intervals based on normality performed as well with a log-normal covariate as with a normally distributed true covariate.

For a logistic regression outcome model, our proposed method gives consistent estimates of the effect of interest when the covariate is normally distributed conditionally on Y_i, i.e. within the diseased and non-diseased groups, as opposed to RC, which is in general inconsistent, although its bias is typically small. For the majority of scenarios we considered, RC had smaller RMSE than conditional normal ML, although for larger sample sizes we would expect conditional normal ML to be superior as bias then dominates the RMSE. Under the more usual assumption of marginal normality for the true covariate, the conditional normal ML estimator had less bias, but larger sampling variability than RC. Despite the fact that the conditional normal ML estimator was misspecified in this situation, there was little bias for the scenarios considered.

Many estimation methods have been proposed which allow some relaxation of parametric assumptions, usually for the unobserved true covariate. At one end of the spectrum are the conditional score and corrected-score approaches [21]. These methods make no assumptions about the distribution of the true covariate, and thus are consistent regardless of the true distribution. They are computationally fast, but are not available in statistical software packages, and are potentially inefficient compared to correctly specified parametric approaches. Non-parametric likelihood methods assume the unobserved covariate is a random variable, but make no further assumptions about its distribution, and thus are also potentially inefficient [22]. In response, methods have been proposed which assume the latent covariate's distribution belongs to a flexible class of continuous distributions [23]. When the outcome model is linear regression, ML based on normality of the unobserved covariate is consistent even when the distribution is not normal, and so may be preferable to these more complex approaches. With a logistic regression outcome model, consistency of our ML implementation relies on the distributional assumptions for the unobserved covariate, and so use of semiparametric approaches such as the conditional score method may be advisable when normality is in doubt.

Recently, parametric multiple imputation (MI) has been proposed and investigated as a method for dealing with covariate measurement error. Cole, Chu and Greenland proposed using MI with a survival outcome model and validation data [24], while Messer and Natarajan considered its implementation for a binary outcome when validation data are available [9]. Freedman et al considered an implementation of MI when measurements may be biased, but a substudy with unbiased replicates is available [16]. With replication data, the true covariate is missing for all study subjects, and so conventional software for MI cannot usually be used. Having fitted the random-intercepts model for $\mathbf{W}_i$ given Y_i in our ML approach, imputations of X_i can be made by drawing values of the random-intercepts, with mean equal to the best linear unbiased prediction (BLUP) and variance equal to its conditional variance [10]. As the number of imputations increases, the resulting MI estimate will converge to the ML estimate, while for a finite number of imputations the MI estimator is less efficient than the ML estimator, and so there may be little benefit in using

imputation in this case [25]. However, MI may be useful when the outcome model is logistic regression and there are error-free covariates which are not jointly normal with the unobserved covariate(s). In such a situation, having fitted the linear mixed model for $\mathbf{W}_i$ given Y_i (and the error-free covariates), one can multiply impute the unobserved X_i from its conditional distribution given observed data, fit the logistic regression model to each of the imputations, and average the resulting parameter estimates across imputations in the usual way.

Our proposal for using ML for covariate measurement error is easily implemented using standard random-intercepts models, which can be fitted using many modern statistical packages. The ML estimate is obtained by simply substituting the relevant parameter estimates into equation (11) for a linear regression outcome model and equation (16) for a logistic regression outcome model. We believe the exposition of the ML approach via random-intercepts models makes the ML approach more transparent, which in turn increases its viability as an alternative to simpler methods such as RC.

A Confidence intervals for $\hat{\beta}_{ML}$

A.1 Linear regression

We restate equation (11) which gives β, the parameter of interest, in terms of the alternative model parametrization:

$$\beta = \frac{\gamma_Y \sigma_Y^2}{\sigma_{X|Y}^2 + \gamma_Y^2 \sigma_Y^2}. \tag{18}$$

Since the likelihood factors into two components, $\hat{\sigma}_Y^2$ is asymptotically uncorrelated with both $\hat{\gamma}_Y$ and $\hat{\sigma}_{X|Y}^2$. Furthermore, a standard result for linear mixed models is that the estimators of fixed effects parameters are asymptotically uncorrelated with the estimators of the variance component parameters [11]. Thus $\hat{\gamma}_Y$ and $\hat{\sigma}_{X|Y}^2$ are asymptotically uncorrelated, and so for large sample sizes $(\hat{\sigma}_Y^2, \hat{\gamma}_Y, \hat{\sigma}_{X|Y}^2) \sim N\left((\sigma_Y^2, \gamma_Y, \sigma_{X|Y}^2), \mathbf{\Sigma} = \text{diag}(\text{Var}(\hat{\sigma}_Y^2), \text{Var}(\hat{\gamma}_Y), \text{Var}(\hat{\sigma}_{X|Y}^2))\right)$. Then by the multivariate delta method it follows that in large samples [14]

$$\hat{\beta} \sim N\left(\beta, \mathbf{J}\mathbf{\Sigma}\mathbf{J}^T\right)$$

where $\mathbf{J}$ denotes the Jacobian matrix of the transformation $(\sigma_Y^2, \gamma_Y, \sigma_{X|Y}^2) \mapsto \beta$:

$$\mathbf{J} = \begin{pmatrix} \frac{\partial \beta}{\partial \sigma_Y^2} & \frac{\partial \beta}{\partial \gamma_Y} & \frac{\partial \beta}{\partial \sigma_{X|Y}^2} \end{pmatrix}.$$

Partial differentiation of equation (18) then gives, after some simplification:

$$\text{Var}(\hat{\beta}) = \frac{\gamma_Y^2 \sigma_{X|Y}^4 \text{Var}(\hat{\sigma}_Y^2) + \sigma_Y^4 (\sigma_{X|Y}^2 - \sigma_Y^2 \gamma_Y^2)^2 \text{Var}(\hat{\gamma}_Y) + \gamma_Y^2 \sigma_Y^4 \text{Var}(\hat{\sigma}_{X|Y}^2)}{(\sigma_{X|Y}^2 + \gamma_Y^2 \sigma_Y^2)^2}.$$

This variance can be estimated by replacing each parameter by its respective estimate. Using the observed information matrix one can estimate the variance of $\hat{\sigma}_Y^2$ by $\frac{2\hat{\sigma}_Y^4}{n}$. A standard error for $\hat{\gamma}_Y$ is given by the linear mixed model output of software packages, and most packages also give a standard error for $\hat{\sigma}_{X|Y}^2$. An approximate 95% confidence interval for $\hat{\beta}$ can be found as $\hat{\beta} \pm 1.96\sqrt{\hat{\text{Var}}(\hat{\beta})}$.

A.2 Logistic regression

From equation (11) the log odds ratio of interest β is given by:

$$\beta = \frac{\gamma_Y}{\sigma^2_{X|Y}}.$$

Analogous to the linear regression case, using the multivariate delta method after partial differentiation of the expression for β, we have:

$$\text{Var}(\hat{\beta}) = \frac{\text{Var}(\hat{\gamma}_Y)}{\sigma^4_{X|Y}} + \frac{\gamma^2_Y \text{Var}(\hat{\sigma}^2_{X|Y})}{\sigma^8_{X|Y}}.$$

Substituting estimates of γ_Y and $\sigma^2_{X|Y}$ and their standard errors gives an estimate of $\text{Var}(\hat{\beta})$, from which a Wald type confidence interval can be calculated.

An alternative confidence interval for $\hat{\beta}$ can be found using Fieller's theorem [26], as the ratio of two normally distributed random variables. Using the fact that $\hat{\gamma}_Y$ and $\hat{\sigma}^2_{X|Y}$ are asymptotically uncorrelated, a 95% confidence interval for $\hat{\beta}$ can be found as:

$$\left(\frac{f_1 - \sqrt{f_1^2 - f_0 f_2}}{f_2}, \frac{f_1 + \sqrt{f_1^2 - f_0 f_2}}{f_2} \right)$$

where

$$\begin{aligned} f_0 &= \hat{\gamma}^2_Y - 1.96^2 \text{Var}(\hat{\gamma}_Y) \\ f_1 &= \hat{\gamma}_Y \hat{\sigma}^2_{X|Y} \\ f_2 &= \hat{\sigma}^4_{X|Y} - 1.96^2 \text{Var}(\hat{\sigma}^2_{X|Y}) \end{aligned}$$

B Multiple covariates

B.1 Linear regression

Let $\mathbf{X}_i = (X_{i1}, .., X_{ip})^T$ denote a p-vector of covariates for subject i which can only be measured with error, and $\mathbf{Z}_i$ denote a q-vector of error-free covariates. To avoid having to specify the distribution of $\mathbf{Z}_i$ specify our model conditional on $\mathbf{Z}_i$. We assume that $\mathbf{X}_i \sim N(\Gamma \mathbf{Z}_i, \Sigma_{X|Z})$, where Γ is a $p \times q$ matrix of regression coefficients. To reduce notational complexity, we assume that first element of $\mathbf{Z}_i$ is equal to 1, representing an intercept term.

Suppose Y_i follows a linear regression model given $\mathbf{X}$ and $\mathbf{Z}$:

$$Y_i = \boldsymbol{\beta}_X^T \mathbf{X}_i + \boldsymbol{\beta}_Z^T \mathbf{Z}_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2_\epsilon)$ is independent of $\mathbf{X}_i$ and $\mathbf{Z}_i$.

Let N_{ij} denote the number of error-prone measurements available for subject i of covariate X_{ij}. As before we assume classical measurement error, so that the kth error-prone measurement of X_{ij}, denoted W_{ijk}, is given by:

$$W_{ijk} = X_{ij} + U_{ijk}$$

where $U_{ijk} \sim N(0, \sigma^2_{U_j})$ is measurement error. We assume the measurement errors U_{ijk} are independent across k, j, and are independent of $\mathbf{X}_i$, $\mathbf{Z}_i$ and ϵ_i. We write $\mathbf{W}_{ij} = (W_{ij1}, .., W_{ijN_{ij}})^T$ for the vector of measurements of covariate j, and $\mathbf{W}_i = (W_{i1}, .., W_{ip})^T$. We can then write

$$\mathbf{W}_i = \mathbf{D}_i \mathbf{X}_i + \mathbf{U}_i$$

where

$$\begin{aligned} \mathbf{D}_i &= \mathbf{1}_{N_{i1}\times 1} \oplus \mathbf{1}_{N_{i2}\times 1} \oplus ... \oplus \mathbf{1}_{N_{ip}\times 1}, \\ \mathbf{U}_i &= (\mathbf{U}_{i1}, .., \mathbf{U}_{ip})^T \end{aligned}$$

and

$$\mathbf{U}_{ij} = (U_{ij1}, .., U_{ijN_{ij}})^T$$

for $j = 1, ..p$.

Analogous to the univariate X case, one can show that $Y_i|Z_i \sim N(\boldsymbol{\delta}^T \mathbf{Z}_i, \sigma^2_{Y|Z})$ where

$$\boldsymbol{\delta}_Y = \Gamma^T \boldsymbol{\beta}_X + \boldsymbol{\beta}_Z \tag{19}$$

$$\sigma^2_{Y|Z} = \boldsymbol{\beta}_X^T \Sigma_X \boldsymbol{\beta}_X + \sigma^2_\epsilon \tag{20}$$

and that $\mathbf{W}_i$ is normal given Y_i with

$$E(\mathbf{W}_i|Y_i) = \mathbf{D}_i(\boldsymbol{\gamma}_Z \mathbf{Z}_i + \gamma_Y Y_i)$$

and

$$\mathrm{Var}(\mathbf{W}_i|Y_i) = \mathbf{D}_i \Sigma_{X|Z,Y} \mathbf{D}_i^T + \Sigma_{\mathbf{U}_i}$$

where

$$\boldsymbol{\gamma}_Z = \Gamma - \Sigma_X \boldsymbol{\beta}_X \sigma^{-2}_{Y|Z} (\boldsymbol{\beta}_X^T \Gamma + \boldsymbol{\beta}_Z^T) \tag{21}$$

$$\gamma_Y = \Sigma_X \boldsymbol{\beta}_X \sigma^{-2}_{Y|Z} \tag{22}$$

$$\Sigma_{X|Z,Y} = \Sigma_X - \Sigma_X \boldsymbol{\beta}_X \sigma^{-2}_{Y|Z} \boldsymbol{\beta}_X^T \Sigma_X \tag{23}$$

$$\Sigma_{\mathbf{U}_i} = \sigma^2_{U_1} \mathbf{I}_{N_{i1}} \oplus \sigma^2_{U_2} \mathbf{I}_{N_{i2}} \oplus ... \oplus \sigma^2_{U_p} \mathbf{I}_{N_{ip}}. \tag{24}$$

This means that $\mathbf{W}_i$ follows a linear mixed model given Y_i and $\mathbf{Z}_i$. The structure of $\mathbf{D}_i$ means that the fixed effects of $\mathbf{Z}_i$ and Y_i are distinct for each of the p covariates contained in $\mathbf{X}_i$. The variance covariance matrix for $\mathbf{W}_i$ given Y_i and $\mathbf{Z}_i$ means that there are p random effects with an unstructured variance covariance matrix $\Sigma_{X|Z,Y}$ with design matrix $\mathbf{D}_i$. Lastly, the residual covariance matrix is the same as for $\mathbf{W}_i$, i.e. a diagonal matrix, with the values along the diagonal corresponding to the measurement error variances of the different measurements.

The likelihood component corresponding to Y_i given $\mathbf{Z}_i$ can be maximized by fitting the least squares regression of Y_i on $\mathbf{Z}_i$, giving the ML estimate of $\boldsymbol{\delta}_Y$. Similarly the residual variance estimate from this regression is an estimate of $\sigma^2_{Y|Z}$. We illustrate how the linear mixed model for $\mathbf{W}_i$ given Y_i and $\mathbf{Z}_i$ can be fitted using SAS Proc Mixed. We assume the dataset is in so called 'long' form, where each row corresponds to a single measurement W_{ijk}. It is convenient to generate a new variable, which here we call 'covariate', indicating which component of $\mathbf{X}_i$ W_{ijk} is

a measurement of, i.e. for a measurement W_{ijk}, 'covariate' would be equal to j. For example, with $p = 2$ covariates measured with error and $\mathbf{Z}_i = (1, Z_i^1)^T$, representing a constant term and a single error-free covariate Z_i^1, the dataset rows corresponding to a subject with one measurement of the first unobserved covariate and two measurements of the second would be:

Y_i	Z_i^1	W_{ijk}	covariate
Y_1	Z_1^1	W_{111}	1
Y_1	Z_1^1	W_{121}	2
Y_1	Z_1^1	W_{122}	2

The linear mixed model could then be specified in SAS by:

```
proc mixed data=myDataset;
  class covariate;
  model w = covariate z1(covariate) y(covariate) / noint;
  random covariate / subject=id type=un;
  repeated / group=covariate;
end;
```

Analogous to the univariate X_i case, one can show that

$$\boldsymbol{\beta}_X = (\Sigma_{X|Z,Y} + \sigma^2_{Y|Z}\boldsymbol{\gamma}_Y\boldsymbol{\gamma}_Y^T)^{-1}\boldsymbol{\gamma}_Y\sigma^2_{Y|Z}$$

and that

$$\boldsymbol{\beta}_Z = \boldsymbol{\delta}_Y - (\boldsymbol{\delta}_Y\boldsymbol{\gamma}_Y^T + \boldsymbol{\gamma}_Z^T)\boldsymbol{\beta}_X.$$

The ML estimates of $\boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_Z$ can be calculated analogously to the univariate X_i case using these formulae.

B.2 Logistic regression

The results for logistic regression can also be extended when $\mathbf{X}_i$ and $\mathbf{Z}_i$ are jointly normal given Y_i. Thus suppose that:

$$\begin{pmatrix}\mathbf{X}_i \\ \mathbf{Z}_i\end{pmatrix} \sim N\left(\begin{pmatrix}\boldsymbol{\gamma}_{X0} + \boldsymbol{\gamma}_{XY}Y_i \\ \boldsymbol{\gamma}_{Z0} + \boldsymbol{\gamma}_{ZY}Y_i\end{pmatrix}, \begin{pmatrix}\Sigma_{X|Y} & \Sigma_{XZ|Y} \\ \Sigma_{ZX|Y} & \Sigma_{Z|Y}\end{pmatrix}\right).$$

This implies that Y_i follows a logistic regression given $\mathbf{X}_i$ and $\mathbf{Z}_i$, with log odds ratios:

$$\begin{pmatrix}\boldsymbol{\beta}_X \\ \boldsymbol{\beta}_Z\end{pmatrix} = \begin{pmatrix}\Sigma_{X|Y} & \Sigma_{XZ|Y} \\ \Sigma_{ZX|Y} & \Sigma_{Z|Y}\end{pmatrix}^{-1}\begin{pmatrix}\gamma_{XY} \\ \gamma_{ZY}\end{pmatrix}. \tag{25}$$

It also follows that $\mathbf{X}_i$ is normal given $\mathbf{Z}_i$ and Y_i, with:

$$\begin{aligned} E(\mathbf{X}_i|\mathbf{Z}_i, Y_i) &= \boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_Y Y_i + \boldsymbol{\gamma}_Z\mathbf{Z}_i \\ \mathrm{Var}(\mathbf{X}_i|\mathbf{Z}_i, Y_i) &= \Sigma_{X|Z,Y} \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\gamma}_0 &= \boldsymbol{\gamma}_{X0} - \Sigma_{XZ|Y}\Sigma^{-1}_{Z|Y}\boldsymbol{\gamma}_{Z0} \\ \boldsymbol{\gamma}_Y &= \boldsymbol{\gamma}_{XY} - \Sigma_{XZ|Y}\Sigma^{-1}_{Z|Y}\boldsymbol{\gamma}_{ZY} \\ \boldsymbol{\gamma}_Z &= \Sigma_{XZ|Y}\Sigma^{-1}_{Z|Y} \\ \Sigma_{X|Z,Y} &= \Sigma_{X|Y} - \Sigma_{XZ|Y}\Sigma^{-1}_{Z|Y}\Sigma_{ZX|Y}. \end{aligned}$$

Now we assume error-prone measurements of $\mathbf{X}_i$ are available as described in the previous section for linear regression. In this case we have:

$$\begin{aligned} E(\mathbf{W}_i|\mathbf{Z}_i, Y_i) &= \mathbf{D}_i\boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_Y Y_i + \boldsymbol{\gamma}_Z \mathbf{Z}_i \\ \mathrm{Var}(\mathbf{W}_i|\mathbf{Z}_i, Y_i) &= \mathbf{D}_i\Sigma_{X|Z,Y}\mathbf{D}_i^T + \Sigma_{U_i}. \end{aligned}$$

This again is a linear mixed model which can be fitted using standard software. In contrast to the previous section, here we have a constant term in the notation in the fixed effects part of the model.

Multivariate regression of $\mathbf{Z}_i$ on Y_i gives the ML estimates of $\boldsymbol{\gamma}_{Z0}$, $\boldsymbol{\gamma}_{ZY}$ and $\Sigma_{Z|Y}$. Fitting the above linear mixed model to $\mathbf{W}_i$ given $\mathbf{Z}_i$ and Y_i then gives the ML estimates of $\boldsymbol{\gamma}_0$, $\boldsymbol{\gamma}_Y$, $\boldsymbol{\gamma}_Z$ and $\Sigma_{X|Z,Y}$. It is then straightforward to show that:

$$\begin{aligned} \boldsymbol{\gamma}_Z &= \Sigma_{XZ|Y}\Sigma_{Z|Y}^{-1} \\ \boldsymbol{\gamma}_{XY} &= \boldsymbol{\gamma}_Y + \boldsymbol{\gamma}_Z\boldsymbol{\gamma}_{ZY} \\ \Sigma_{XZ|Y} &= \boldsymbol{\gamma}_Z\Sigma_{Z|Y} \\ \Sigma_{X|Y} &= \Sigma_{X|Z,Y} + \Sigma_{XZ|Y}\Sigma_{Z|Y}^{-1}\Sigma_{ZX|Y}. \end{aligned}$$

These formulae can be used to calculate the ML estimates of these parameters, and then these can be inserted into equation (25) to calculate the ML estimates of $\boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_Z$.

C Unbiasedness of likelihood score equations for the mixed model for $\mathbf{W}_i$ given Y_i

We show that the likelihood score equations for the linear mixed model for $\mathbf{W}_i$ given Y_i remain unbiased without requiring normality of X_i or U_{ij}. Irrespective of whether X_i and ϵ_i are both normally distributed, we can always express X_i in terms of its best linear prediction given Y_i as:

$$X_i = \gamma_0 + \gamma_Y Y_i + b_i \tag{26}$$

where $E(b_i) = 0, \mathrm{Cov}(Y_i, b_i) = 0$, and γ_0 and γ_Y are as given in equations (8) and (9). The random term b_i is no longer necessarily normally distributed, and its variance may vary as a function of Y_i. The expression in equation (10) for $\sigma^2_{X|Y}$ then equals the mean conditional variance of the b_i. Thus Y_i and b_i are not necessarily independent, although they are uncorrelated. We can write:

$$\begin{aligned} \mathbf{W}_i &= X_i\mathbf{1}_{N_i\times 1} + \mathbf{U}_i \\ &= \gamma_0\mathbf{1}_{N_i\times 1} + \gamma_Y Y_i\mathbf{1}_{N_i\times 1} + b_i\mathbf{1}_{N_i\times 1} + \mathbf{U}_i. \end{aligned} \tag{27}$$

We now let $\boldsymbol{\gamma} = (\gamma_0, \gamma_Y)^T$ and let $\mathcal{X}_i$ denote the $N_i \times 2$ design matrix for the fixed effects in the above model. This is simply a matrix whose first column has entries equal to one and second column with entries Y_i. In this notation:

$$\mathbf{W}_i = \mathcal{X}_i\boldsymbol{\gamma} + b_i\mathbf{1}_{N_i\times 1} + \mathbf{U}_i. \tag{28}$$

From standard linear mixed model theory (for example Chapter 6 of [11]), the likelihood score corresponding to the fixed effects of the linear mixed model for $\mathbf{W}_i$ given Y_i can be expressed as:

$$\mathcal{X}_i^T\mathbf{V}_i^{-1}\mathbf{W}_i - \mathcal{X}_i^T\mathbf{V}_i^{-1}\mathcal{X}_i\boldsymbol{\gamma} \tag{29}$$

where $\mathbf{V}_i = \sigma^2_{X|Y}\mathbf{1}_{N_i\times N_i} + \sigma^2_U\mathbf{I}_{N_i}$ denotes the variance covariance matrix of $\mathbf{W}_i$ given Y_i under the assumed model. We now substitute for $\mathbf{W}_i$ using equation (28), which gives

$$\begin{aligned}\mathcal{X}_i^T\mathbf{V}_i^{-1}(\mathcal{X}_i\boldsymbol{\gamma} + b_i\mathbf{1}_{N_i\times 1} + \mathbf{U}_i) - \mathcal{X}_i^T\mathbf{V}_i^{-1}\mathcal{X}_i\boldsymbol{\gamma} \\ = \mathcal{X}_i^T\mathbf{V}_i^{-1}\mathbf{1}_{N_i\times 1}b_i + \mathcal{X}_i^T\mathbf{V}_i^{-1}\mathbf{U}_i.\end{aligned} \tag{30}$$

Now since $E(b_i) = 0$, taking expectations (conditional on N_i) gives:

$$E(\mathcal{X}_i^T\mathbf{V}_i^{-1}\mathbf{1}_{N_i\times 1}b_i) = \mathrm{Cov}(\mathcal{X}_i^T\mathbf{V}_i^{-1}\mathbf{1}_{1\times N_i}, b_i) = \mathbf{0}$$

because $\mathrm{Cov}(Y_i, b_i) = 0$. Similarly, because $\mathrm{Cov}(Y_i, U_{ij}) = 0$ (because of the assumed independence between U_{ij} and both X_i and ϵ_i), $E(\mathcal{X}_i^T\mathbf{V}_i^{-1}\mathbf{U}_i) = 0$.

The likelihood score corresponding to $\sigma^2_{X|Y}$ is given by:

$$-0.5\mathrm{tr}(\mathbf{V}_i^{-1}\mathbf{1}_{N_i\times N_i}) + 0.5(\mathbf{W}_i - \mathcal{X}_i\boldsymbol{\gamma})^T\mathbf{V}_i^{-1}\mathbf{1}_{N_i\times N_i}\mathbf{V}_i^{-1}(\mathbf{W}_i - \mathcal{X}_i\boldsymbol{\gamma}) \tag{31}$$

where tr() denotes the trace of a square matrix. By a standard result for the expectation of quadratic forms (e.g. Appendix S5 of [11]), the expectation of the expectation of the second part of the likelihood score is equal to:

$$0.5(E(\mathbf{W}_i - \mathcal{X}_i\boldsymbol{\gamma})^T\mathbf{V}_i^{-1}\mathbf{1}_{N_i\times N_i}\mathbf{V}_i^{-1}E(\mathbf{W}_i - \mathcal{X}_i\boldsymbol{\gamma}) + \mathrm{tr}(\mathbf{V}_i^{-1}\mathbf{1}_{N_i\times N_i}\mathbf{V}_i^{-1}\mathrm{Var}(\mathbf{W}_i - \mathcal{X}_i\boldsymbol{\gamma}))). \tag{32}$$

Then since $\mathbf{W}_i - \mathcal{X}_i\boldsymbol{\gamma} = \mathbf{1}_{N_i\times 1}b_i + \mathbf{U}_i$, it follows that $E(\mathbf{W}_i - \mathcal{X}_i\boldsymbol{\gamma}) = \mathbf{0}$. Furthermore, the assumption of independence between the errors U_{ij} and both X_i and ϵ_i imply that b_i is uncorrelated with U_{ij}. This in turn means that $\mathrm{Var}(\mathbf{W}_i - \mathcal{X}_i\boldsymbol{\gamma}) = \mathbf{V}_i$, and so the expectation of the second part of the score is equal to:

$$0.5\mathrm{tr}(\mathbf{V}_i^{-1}\mathbf{1}_{N_i\times N_i}\mathbf{V}_i^{-1}\mathbf{V}_i) \quad = 0.5\mathrm{tr}(\mathbf{V}_i^{-1}\mathbf{1}_{N_i\times N_i}). \tag{33}$$

The expectation of the likelihood score corresponding to $\sigma^2_{X|Y}$ thus has expectation zero.

Lastly we consider the likelihood score component corresponding to σ^2_U, which is given by:

$$-0.5\mathrm{tr}(\mathbf{V}_i^{-1}) + 0.5(\mathbf{W}_i - \mathcal{X}_i\boldsymbol{\gamma})^T\mathbf{V}_i^{-1}\mathbf{V}_i^{-1}(\mathbf{W}_i - \mathcal{X}_i\boldsymbol{\gamma}). \tag{34}$$

Taking expectations, and as above using the fact that $E(\mathbf{W}_i - \mathcal{X}_i\boldsymbol{\gamma}) = \mathbf{0}$, that $\mathrm{Var}(\mathbf{W}_i - \mathcal{X}_i\boldsymbol{\gamma}) = \mathbf{V}_i$, and the rule for the expectation of quadratic forms, we have:

$$-0.5\mathrm{tr}(\mathbf{V}_i^{-1}) + 0.5\mathrm{tr}(\mathbf{V}_i^{-1}\mathbf{V}_i^{-1}\mathbf{V}_i) = 0 \tag{35}$$

so that the likelihood score for σ^2_U also has expectation zero. Since the conditional expectation of the score equations given N_i is zero, the marginal expectation is also zero.

D Acknowledgements

This research was supported by the Medical Research Council in the UK, through a PhD studentship award to the first author. We thank the anonymous reviewer for their suggestions, which improved the manuscript.

References

[1] Fuller WA. *Measurement Error Models*. Wiley: New York, 1987.

[2] Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement Error in Nonlinear Models* (2nd edn.). Chapman & Hall / CRC: Boca Raton, 2006.

[3] Kuha J. Corrections for exposure measurement error in logistic regression models with an application to nutritional data. *Statistics in Medicine* 1994; **13**:1135-1148. DOI: 10.1002/sim.4780131105

[4] Schafer DW, Purdy KG. Likelihood analysis for errors-in-variables regression with replicate measurements. *Biometrika* 1996; **83**:813-824. DOI: 10.1093/biomet/83.4.813

[5] Rabe-Hesketh S, Skrondal A, Pickles A. Maximum likelihood estimation of generalized linear models with covariate measurement error. *The Stata Journal* 2003; **3**:386-411.

[6] Skrondal A, Rabe-Hesketh S. *Generalized Latent Variable Modelling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall / CRC: Boca Raton, 2004.

[7] Clayton D. Models for the analysis of cohort and case-control studies with inaccurately measured exposures. In Dwyer JH, Feinleib M, Lipsert P, Hoffmeister H, (Eds), *Statistical Models for Longitudinal Studies of Health*. Oxford University Press, New York, 1991.

[8] Muthén LK & Muthén BO. *Mplus User's Guide*. Muthén & Muthén, Los Angeles, California, 1998.

[9] Messer K, Natarajan L. Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment. *Statistics in Medicine* 2008; **27**:6332-6350. DOI: 10.1002/sim.3458

[10] Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*. Springer: New York, 2000.

[11] Searle SR, Casella G, McCulloch CE. *Variance Components*. Wiley: New York, 1992.

[12] Lindsey JK. *Parametric Statistical Inference*. Oxford University Press: New York, 1996.

[13] R Development Core Team. *R: A Language and Environment for Statistical Computing*. http://www.R-project.org R Foundation for Statistical Computing, Vienna, Austria, 2008.

[14] Casella G, Berger RL. *Statistical Inference*. Duxbury Press, Pacific Grove, California, 2001.

[15] Freedman LS, Fainberg V, Kipnis D, Midthune D, Carroll RJ. A new method for dealing with measurement error in explanatory variables of regression models. *Biometrics* 2004; **60**:172-181. DOI: 10.1111/j.0006-341X.2004.00164.x

[16] Freedman LS, Midthune D, Carroll RJ, Kipnis V. A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. *Statistics in Medicine* 2008; **27**:5195-5216. DOI: 10.1002/sim.3361

[17] Gleser LJ. Improvements in the naive approach to estimation in nonlinear errors-in-variables regression models. In Brown PJ, Fuller WA, (eds), *Statistical Analysis of Measurement Error Models and Applications*. American Mathematical Society, Providence, 1990.

[18] Westfall PH. A comparison of variance component estimates for arbitrary underlying distributions. *Journal of the American Statistical Association* 1987; **82**:866-874.

[19] Efron B. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association* 1975; **70**892-898.

[20] Press SJ, Wilson S. Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association* 1978; **73**699-705.

[21] Stefanski LA, Carroll RJ. Conditional scores and optimal scores for generalized linear measurement error models. *Biometrika* 1987; **74**:703:716. DOI: 10.1093/biomet/74.4.703

[22] Rabe-Hesketh S, Pickles A, Skrondal A. Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statistical Modelling* 2003; **3**:215-232. DOI: 10.1191/1471082X03st056oa

[23] Li E, Zhang D, Davidian M. Likelihood and pseudo-likelihood methods for semiparametric joint models for a primary endpoint and longitudinal data. *Computational Statistics & Data Analysis* 2007; **51**:5776-5790. DOI: 10.1016/j.csda.2006.10.008

[24] Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction *International Journal of Epidemiology* 2006; **35**:1074-1081. DOI: 10.1093/ije/dyl097

[25] Wang N, Robins JM. Large-sample theory for parametric multiple imputation procedures. *Biometrika* 1998; **85**:935-948. DOI: 10.1093/biomet/85.4.935

[26] Buonaccorsi JP. Fieller's theorem. In Armitage P, Colton T (eds), *Encyclopedia of Biostatistics*. Wiley, Chichester, 1998.

Table 1: Simulation results for $\hat{\beta}$ using regression calibration (RC) and maximum likelihood (ML), based on 10,000 simulations in which Y follows a linear regression model given X with $\beta = 1$ with normally distributed residuals and normally distributed X; n=5,000 subjects, 4,500 with one error-prone measurement of X, 500 with two

			RC	ML	
Scenario	corr(Y, X)	λ	Mean (SD)	Mean (SD)	CI coverage*
1	0.2	2/3	1.001 (0.092)	1.001 (0.092)	94.9 (97.6, 97.3)
2	0.2	1/2	1.004 (0.116)	1.004 (0.116)	95.1 (98.6, 96.5)
3	0.2	1/3	1.014 (0.171)	1.014 (0.171)	95.1 (99.5, 95.6)
4	0.5	2/3	1.000 (0.043)	1.000 (0.043)	95.3 (98.5, 96.8)
5	0.5	1/2	1.003 (0.070)	1.003 (0.069)	94.7 (99.1, 95.7)
6	0.5	1/3	1.017 (0.129)	1.017 (0.125)	95.2 (99.8, 95.4)
7	0.8	2/3	1.001 (0.034)	1.001 (0.033)	95.0 (98.7, 96.2)
8	0.8	1/2	1.004 (0.063)	1.003 (0.057)	95.1 (99.0, 96.1)
9	0.8	1/3	1.013 (0.121)	1.011 (0.108)	94.8 (99.8, 95.0)

λ - ratio of variance of X to variance of error-prone measurements
SD - standard deviation
* empirical coverage of 95% Wald-type confidence intervals (β_L, β_U) (coverage of one-sided 97.5% intervals (β_L, ∞), coverage of one-sided 97.5% intervals $(-\infty, \beta_U)$)

Table 2: Simulation results for $\hat{\beta}$ using regression calibration (RC) and maximum likelihood (ML), based on 10,000 simulations in which Y follows a linear regression model given X with $\beta = 1$ with normally distributed residuals and log-normally distributed X; n=5,000 subjects, 4,500 with one error-prone measurement of X, 500 with two

			RC	ML	
Scenario	corr(Y, X)	λ	Mean (SD)	Mean (SD)	CI coverage*
1	0.2	2/3	1.000 (0.091)	1.004 (0.091)	95.0 (97.5, 97.5)
2	0.2	1/2	1.000 (0.122)	1.008 (0.122)	94.4 (98.0, 96.4)
3	0.2	1/3	1.010 (0.185)	1.022 (0.186)	94.1 (99.4, 94.8)
4	0.5	2/3	0.999 (0.047)	1.002 (0.046)	94.1 (97.8, 96.3)
5	0.5	1/2	1.001 (0.079)	1.007 (0.076)	93.6 (98.5, 95.0)
6	0.5	1/3	1.012 (0.145)	1.018 (0.134)	93.7 (99.8, 94.0)
7	0.8	2/3	0.998 (0.038)	1.001 (0.034)	94.0 (98.2, 95.8)
8	0.8	1/2	1.000 (0.073)	1.004 (0.060)	94.3 (98.8, 95.5)
9	0.8	1/3	1.014 (0.146)	1.014 (0.112)	94.2 (99.6, 94.6)

λ - ratio of variance of X to variance of error-prone measurements

SD - standard deviation

* empirical coverage of 95% Wald-type confidence intervals (β_L, β_U) (coverage of one-sided 97.5% intervals (β_L, ∞), coverage of one-sided 97.5% intervals $(-\infty, \beta_U)$)

Table 3: Simulation results for $\hat{\beta}$ using regression calibration (RC) and maximum likelihood (ML), based on 10,000 simulations in which Y follows a logistic regression model given X with log odds ratio β, $X|Y$ normally distributed; n=5,000 subjects, 4,500 with one error-prone measurement of X, 500 with two

				RC		ML			
Scenario	$P(Y=1)$	β	λ	Mean (SD)	RMSE	Mean (SD)	RMSE	Wald CI coverage*	Fieller CI coverage*
1	0.1	0.1	2/3	0.100 (0.058)	0.058	0.100 (0.058)	0.058	95.0 (97.6, 97.4)	95.0 (97.5, 97.5)
2			1/2	0.101 (0.066)	0.066	0.101 (0.066)	0.066	95.7 (98.0, 97.7)	95.5 (97.7, 97.8)
3			1/3	0.102 (0.083)	0.083	0.102 (0.084)	0.084	95.7 (98.2, 97.5)	95.0 (97.5, 97.6)
4		1	2/3	0.976 (0.070)	0.074	1.003 (0.075)	0.075	95.1 (97.8, 97.3)	95.0 (97.2, 97.8)
5			1/2	0.966 (0.092)	0.099	1.007 (0.103)	0.104	95.2 (98.8, 96.4)	94.8 (97.3, 97.5)
6			1/3	0.961 (0.140)	0.145	1.019 (0.162)	0.163	95.1 (99.9, 95.3)	95.0 (97.7, 97.3)
7	0.5	0.1	2/3	0.100 (0.035)	0.035	0.100 (0.035)	0.035	95.0 (97.7, 97.3)	95.0 (97.5, 97.4)
8			1/2	0.100 (0.041)	0.041	0.100 (0.041)	0.041	95.3 (97.9, 97.4)	95.0 (97.5, 97.6)
9			1/3	0.101 (0.051)	0.051	0.102 (0.051)	0.051	95.9 (98.3, 97.6)	95.3 (97.4, 97.9)
10		1	2/3	0.940 (0.052)	0.079	1.003 (0.063)	0.063	95.1 (98.4, 96.7)	95.0 (97.5, 97.5)
11			1/2	0.915 (0.073)	0.113	1.007 (0.096)	0.096	95.0 (99.3, 95.7)	94.8 (97.8, 97.1)
12			1/3	0.897 (0.122)	0.160	1.024 (0.170)	0.172	94.1 (100, 94.1)	94.6 (97.6, 97.0)

λ - ratio of variance of X to variance of error-prone measurements

SD - standard deviation

RMSE - root mean squared error

* empirical coverage of 95% Wald/Fieller confidence intervals (β_L, β_U) (coverage of one-sided 97.5% intervals (β_L, ∞), coverage of one-sided 97.5% intervals $(-\infty, \beta_U)$)

Table 4: Simulation results for $\hat{\beta}$ using regression calibration (RC) and maximum likelihood (ML), based on 10,000 simulations in which Y follows a logistic regression model given X with log odds ratio β, X normally distributed; n=5,000 subjects, 4,500 with one error-prone measurement of X, 500 with two

				RC		ML			
Scenario	$P(Y=1)$	β	λ	Mean (SD)	RMSE	Mean (SD)	RMSE	Wald CI coverage*	Fieller CI coverage*
1	0.1	0.1	2/3	0.100 (0.057)	0.057	0.100 (0.058)	0.058	95.1 (97.6, 97.4)	95.0 (97.6, 97.5)
2			1/2	0.100 (0.067)	0.067	0.101 (0.067)	0.067	95.1 (97.7, 97.4)	94.9 (97.5, 97.4)
3			1/3	0.102 (0.082)	0.082	0.102 (0.082)	0.082	95.6 (98.1, 97.6)	95.1 (97.4, 97.6)
4		1	2/3	0.966 (0.069)	0.077	0.982 (0.073)	0.075	93.9 (99.2, 94.8)	94.5 (98.8, 95.6)
5			1/2	0.954 (0.091)	0.102	0.987 (0.100)	0.101	94.2 (99.3, 94.9)	94.8 (98.5, 96.3)
6			1/3	0.946 (0.140)	0.150	0.997 (0.160)	0.160	93.2 (99.9, 93.3)	94.2 (98.2, 96.0)
7	0.5	0.1	2/3	0.100 (0.035)	0.035	0.100 (0.035)	0.035	94.7 (97.6, 97.1)	94.6 (97.4, 97.2)
8			1/2	0.101 (0.040)	0.040	0.101 (0.040)	0.040	95.4 (97.8, 97.5)	95.3 (97.6, 97.7)
9			1/3	0.101 (0.050)	0.050	0.101 (0.051)	0.051	95.9 (98.6, 97.5)	95.1 (97.5, 97.6)
10		1	2/3	0.938 (0.051)	0.080	1.000 (0.062)	0.062	94.9 (98.6, 96.3)	94.8 (97.7, 97.1)
11			1/2	0.911 (0.071)	0.114	1.002 (0.092)	0.092	94.8 (99.3, 95.5)	94.8 (97.8, 97.0)
12			1/3	0.894 (0.117)	0.158	1.020 (0.163)	0.164	94.8 (100, 94.8)	94.8 (97.8, 97.0)

λ - ratio of variance of X to variance of error-prone measurements

SD - standard deviation

RMSE - root mean squared error

* empirical coverage of 95% Wald/Fieller confidence intervals (β_L, β_U) (coverage of one-sided 97.5% intervals (β_L, ∞), coverage of one-sided 97.5% intervals $(-\infty, \beta_U)$)